DOCUMENT RESUME

ED 135 834 TM 006 059

AUTHOR Scheetz, James P.
TITLE The Use of PBIB Designs in Educational Evaluation.
PUB DATE [Sep 76]
NOTE 19p.; Paper presented at the Annual Meeting of the American Psychological Association (84th, Washington, D.C., September 3-7, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Item Sampling; Statistical Analysis; Testing Programs
IDENTIFIERS *Partially Balanced Incomplete Block Designs

ABSTRACT

When performing large scale evaluations (e.g., on a state-wide or national level) it may not be possible to administer all items in the item universe to all respondents in the subject population. One method which has been proposed to sample both items and respondents is multiple matrix sampling (MMS) in which a sample of the items is administered to a sample of subjects. In this study, the rules for constructing partially balanced incomplete block (PBIB) designs were followed in assigned items to subtests. Estimates of the population mean and variance were obtained and the results were compared with the estimates of the mean and variance obtained from using simple random sampling to construct subtests. In all comparisons between the two methods of assigning items to subtests, the number of subtests, the number of items per subtest and the total number of observations was held constant. The findings suggest that use of the PBIB designs may result in slightly less variable estimates of the mean, but simple random sampling will result in slightly less variable estimates of the variance. The primary conclusion is that the gain in precision which may occur when using PBIB designs is probably not worth the added effort and expense. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED135834

THE USE OF PBIB DESIGNS IN EDUCATIONAL EVALUATION

James P. Scheetz
Health Sciences Center
TEAM Program
University of Louisville
Louisville, Kentucky 40201

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

American Psychological Association
84th Annual Convention
September 3-7, 1976
Washington, D.C.

TM006 059

The Use of PBIB Designs in Educational Evaluation

Introduction

The use of partially balanced incomplete block (PBIB) designs as a means of conducting educational evaluations has been stressed by several authorities, e.g., Shoemaker (1973b), Knapp (1968, 1972). PBIB designs[1] are most appropriately used as a program evaluation tool in situations where multiple matrix sampling (MMS) could be used. Shoemaker (1973b) listed the following as advantages of MMS: (1) reduced standard errors of estimate, (2) a means of testing large item universes, and (3) reduction in testing time. Since PBIB designs may be used to assign items to subtests when using MMS procedures, the same advantages accrue to PBIB designs.

As illustrated in Figure 1, MMS involves the administration of a sample of items drawn from a larger universe of items to a sample of respondents drawn from a larger population.

---

[1] For a complete listing of PBIB designs, see: W. H. Clatworthy, Tables of two-associate-class partially balanced designs. United States Department of Commerce, National Bureau of Standards (Applied Mathematics Series, Volume 63), 1973.

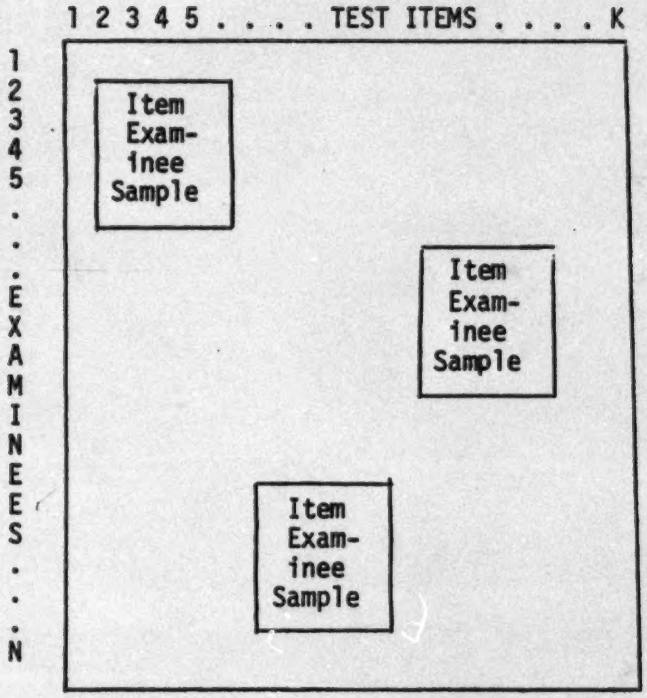

Figure 1

Illustration of Multiple Matrix Sampling

The central question when using MMS procedures is to choose the method of allocating items to subtests which most accurately estimates the parameters of interest, usually the population mean and variance designated by $\hat{\mu}$ and $\hat{\sigma}^2$, respectively. Shoemaker (1970a, 1970b, 1971, 1972, 1973a) has extensively investigated one aspect of this question which deals with the methods of constructing subtests in terms of varying t, k, and n where

t = the number of subtests

k = the number of items per subtest, and

n = the number of respondents per subtest

In addition, three other parameters are of interest. The first is K, the total number of items in the universe, which is equal to the product (t)(k). The second is N, the total number of examinees in the population, and the third is the total number of observations, Z, which is equal to the product (t)(k)(n), such that Z = tkn.

In addition to choosing optimal values of t, k, and n, another part of the allocation question focuses on the sampling procedure used to assign items to a particular subtest.

Several methods can be used to allocate items to a subtest, the simplest being random sampling of items. The primary advantage of using this technique is the ease of implementation. A second approach is to stratify the items by either difficulty level or content area and then use stratified random sampling to allocate items to subtests. The advantage is that within strata variance should be decreased, resulting in a reduction of total variance. A third method of allocating items is to use a PBIB design. The purpose of this paper is to compare the relative efficiency of simple random sampling versus the use of PBIB designs in allocating items to subtests.

In contrast to the typical MMS designs using random or stratified random sampling, a PBIB design is one in which the same item appears on more than one subtest and particular pairs of items occur more frequently than other pairs. The reason for trying to include as many different pairs of items as possible among the subtests is to estimate the total test variance as accurately as possible. The total test variance is equal to the sum of the item variances plus twice the sum of the item covariances as given by the following formula:

$$S_t^2 = \sum_{i=1}^{k} S_i^2 + 2\sum_{\substack{i,j=1 \\ i<j}}^{k} \text{cov}\,(i,j) \qquad (1)$$

where S_t^2 = the total test variance

S_i^2 = the variance of item i

$(i = 1, 2, 3, \ldots, k)$

cov (i,j) = covariance of items i and j

$(i,j) = 1, 2, 3, \ldots k)$

If a particular covariance term is missing because a given item pairing did not occur, this term must be estimated. One common solution to this problem is to use the average of all covariance terms which were computed as the estimate of the missing terms. This method, however, is less satisfactory than actually computing the terms. For example, consider a test of 36 items which will be used in a MMS situation with t = 6 and k = 6. The number of covariance terms is given by [K (K - 1)]/2, a total of 630 covariance terms for a 36 item test. If an item appears on only one subtest, only 90 of these terms are computed.

Description of PBIB Designs Used in this Study

To illustrate the concept of a PBIB design, a very simple example will be used.

	SUBTEST (t)			
	1	2	3	4
ITEMS (k)	4	1	2	1
	5	3	3	2
	6	5	4	6

Figure 2

Illustration of Partially Balanced Incomplete Block Design

In Figure 2, the total number of items in the universe, K, is equal to six, and r = 2 which specifies the number of times each item is repeated in the design. The number of subtests, t, is equal to four. The number of items per subtest, denoted by k, is three. The number of times an item is paired with each first associate is λ_1 and n_1 is the number of first associates. First associates are those items which occur together λ_1 times among the subtests of the design, and all first associates are paired an equal number of times among the subtests. In this example, $\lambda_1 = 1$ and $n_1 = 4$. For purposes of illustration consider item 4 which appears in subtests one and three. Item 4 is paired once each with items 5, 6, 2, and 3; hence $\lambda_1 = 1$ and $n_1 = 4$. The frequency with which an item is paired with each second associate is λ_2 with n_2 being the number of second associates. Second associates are those items which occur together λ_2 times among the subtests of the design. For PBIB designs $\lambda_1 \neq \lambda_2$; if $\lambda_1 = \lambda_2$, the design is a balanced incomplete block (BIB) design. In this example $\lambda_2 = 0$ and $n_2 = 1$, which indicates that item 4 is not paired with one item and this item is item 1.

The parameters of the PBIB designs used in this study are listed

in Table 1.

Table 1

Parameters of the PBIB Designs

Plan	t	k	n	r	Z	λ_1	n_1	λ_2	n_2	NREPS[a]
1	9	8	10	2	720	1	14	0	21	10
2	12	6	10	2	720	1	10	0	25	10
3	36	5	4	5	720	1	20	0	15	10
4	36	10	2	10	720	4	10	0	25	10
5	120	3	2	10	720	2	10	0	25	10

[a] NREPS = the number of times each sampling plan was replicated in the computer analysis.

Each replication of a sampling plan resulted in t estimates of $\hat{\mu}$ and $\hat{\sigma}^2$ (i.e., each subtest resulted in one estimate of $\hat{\mu}$ and $\hat{\sigma}^2$). The standard error terms and the average deviations reported for each sampling plan were based on (t)(NREPS) estimates of $\hat{\mu}$ and $\hat{\sigma}^2$.

Description of Simple Random Sampling Plans Used in this Study

Five designs employing simple random sampling to allocate items to subtests were used and the parameters of these designs are listed in Table 2.

Table 2

Parameters of the Simple Randomized Designs

Plan	t	k	n	Z	NREPS
1	3	12	20	720	10
2	4	9	20	720	10
3	6	6	20	720	10
4	9	4	20	720	10
5	12	3	20	720	10

An examination of Tables 1 and 2 shows that for the PBIB designs, tk exceeded K; for the simple randomized designs, tk = K. This relationship underlies the fundamental difference between the two types of designs: for PBIB designs, items may appear on more than one subtest, whereas the randomized design involves single exhaustion of the item universe with a particular item appearing on only one subtest.

From Tables 1 and 2 it can also be seen that Z = 720 for all sampling plans. This value was held constant because of the importance of the total number of observations in estimating the standard errors of $\hat{\mu}$ and $\hat{\sigma}^2$. For a summary of the findings with respect to Z, see Shoemaker (1972).

In this study, the two types of sampling plans were compared on the basis of: (1) the standard error of $\hat{\mu}$, (2) the standard error of $\hat{\sigma}^2$,

(3) the average absolute deviation of the mean estimated by MMS from the mean of the data base, and (4) the average absolute deviation of the variance estimated by MMS from the variance of the data base. The sampling plans which minimize standard errors of $\hat{\mu}$ and $\hat{\sigma}^2$ and also minimize absolute deviations are superior to those which result in larger standard errors of $\hat{\mu}$ and $\hat{\sigma}^2$ and larger absolute deviations. Therefore, the purpose of this study was to compare the sampling plans and ascertain which type both resulted in smaller standard errors of $\hat{\mu}$ and $\hat{\sigma}^2$ and produced smaller absolute deviations.

Description of the Data Base

The items used in this study were the 36 items of the mathematics subtest of the Iowa Tests of Educational Development. During 1971, 16,819 ninth-grade students took the math test (Form X-6). A 1 in T systematic sample was drawn to give a data base with N = 600. For this data base of 600 the following values were found: $\mu = 11.623$, $\sigma^2 = 30.385$, $\sigma = 5.512$, skewness = .865, and kurtosis = 3.680. The range was 32 with a minimum score of one and a maximum score of 33. The reliability (KR 20) for the data base was .779. The distribution was positively skewed and leptokurtic.

Methods of Comparison

Each replication of a sampling plan resulted in a pooled estimate of the mean and variance. The pooled estimate of any parameter is given by Shoemaker (1973b, p. 38) as

$$\hat{\zeta} = \frac{\sum_{s=1}^{t} Z_s \hat{\zeta}_s}{\sum_{s=1}^{t} Z_s} \qquad (2)$$

where $\hat{\zeta}$ refers to any parameter, t is the number of subtests and $Z_s = n_s k_s$, the number of observations obtained from subtest s. After these estimates were computed, the standard errors (SE) of $\hat{\mu}$ and $\hat{\sigma}^2$ were calculated for each sampling plan. The plans were then ranked from least to greatest in terms of $SE(\hat{\mu})$ and $SE(\hat{\sigma}^2)$ and an average rank for the two types of sampling plans was the basis for comparison.

Two other statistics, Ψ_1 and Ψ_2, involving the average absolute deviations of $\hat{\mu}$ and $\hat{\sigma}^2$ from μ and σ^2 of the data base were calculated in order to compare the sampling plans. For the mean,

$$\Psi_1 = \frac{\Sigma|\mu - \hat{\mu}_{p_i}|}{NREPS} \qquad (3)$$

was computed where μ is the mean for the data base and $\hat{\mu}_{p_i}$ is the estimated mean score for one replication of a given sampling plan. This statistic is of interest because it indicates how accurately the mean computed by the MMS procedure estimates the true mean of the data base. A similar statistic, Ψ_2, was computed to measure the degree of accuracy of the estimated variance. The statistic is

$$\Psi_2 = \frac{\Sigma|\sigma^2 - \hat{\sigma}^2_{p_i}|}{NREPS} \qquad (4)$$

where σ^2 is the variance for the data base and $\hat{\sigma}^2_{p_i}$ is the estimated variance for one replication of a sampling plan.

Results and Discussion

The results of the sampling plans analyzed in the study are presented in Table 3. The first five plans are PBIB designs while the last five involve simple random sampling to assign items to subtests. The method of analysis involved a comparison of the average rank of the PBIB and simple random sampling plans and these results are summarized in Table 4.

The results of the average ranking indicate that when estimating μ, PBIB designs are slightly superior to simple random sampling as a means of assigning items to subtests. The average rank of the plans based on PBIB designs is 5.20 whereas the plans which used random sampling have a rank of 5.80 when $SE(\hat{\mu})$ is the basis of comparison. When compared on the basis of Ψ_1, the rank of the PBIB plans is 5.00 compared with 6.00 for the simple randomized plans.

The results for the variance indicate that the use of PBIB designs resulted in poorer estimates than those found for the simple randomized designs. The average rank with $SE(\hat{\sigma}^2)$ as the criterion is 6.00 for the PBIB plans and 5.00 for the simple randomized plans. Values of 6.00 for the PBIB plans and 5.00 for the simple randomized plans were found with Ψ_2 as the criterion.

The results of the current study partially confirm the findings reported by Shoemaker (1973a). He found that, in general, the PBIB designs produced smaller standard error terms for all moments, but

the magnitude of the discrepancy between the standard error terms for the two types of designs was greatest for the mean and decreased quite markedly for the higher moments. In the present study, the values of $SE(\hat{\mu})$ tend to be less for the PBIB designs, but the $SE(\hat{\sigma}^2)$ tend to be less for the simple randomized designs. The results of this study confirm Shoemaker's findings with regard to the mean, but contradict his conclusions regarding the variance. However, the current findings may be explained by Shoemaker's observation that the difference between the estimates provided by the two types of plans is much greater for the mean than the difference between PBIB estimates and those based on simple random sampling for the higher moments.

TABLE 3

Results of Using PBIB Designs to Assign Items to Subtests - Mathematics

PLAN						POOLED $\hat{\mu}$	SE $\hat{\mu}$	RANK	Ψ_1	RANK	POOLED $\hat{\sigma}^2$	SE $\hat{\sigma}^2$	RANK	Ψ_2	RANK
	t	k	n	Z	NREPS										
	120	3	2	720	10										
1	λ_1	n_1	λ_2	n_2	r	11.620	.546	5	.405	3	32.201	10.420	8	8.813	7
	2	10	0	25	10										
	t	k	n	Z	NREPS										
	36	10	2	720	10										
2	λ_1	n_1	λ_2	n_2	r	11.910	.635	8	.546	8	31.169	9.383	6	8.100	6
	4	10	2	25	10										
	t	k	n	Z	NREPS										
	36	5	4	720	10										
3	λ_1	n_1	λ_2	n_2	r	11.665	.683	9	.555	9	30.594	6.658	3	5.415	3
	0	15	1	20	5										
	t	k	n	Z	NREPS										
	12	6	10	720	10										
4	λ_1	n_1	λ_2	n_2	r	11.745	.483	1	.390	1	26.760	10.905	9	9.510	9
	1	10	0	25	2										

TABLE 3 Cont.

	t	k	n	Z	NREPS									
	9	8	10	720	10									
5	λ_1	n_1	λ_2	n_2	r	11.550	.534	3	.425	4	24.444	6.811	4	7.349	5
	1	14	0	21	2										

	t	k	n	Z	NREPS									
6	12	3	20	720	10	11.620	.615	7	.399	2	29.257	9.342	5	7.346	4
7	9	4	20	720	10	11.320	.586	6	.489	7	31.066	5.512	2	4.239	2
8	6	6	20	720	10	11.495	.542	4	.465	6	30.735	11.858	10	9.717	10
9	4	9	20	720	10	11.595	.498	2	.450	5	28.150	5.147	1	4.073	1
10	3	12	20	720	10	11.745	.691	10	.576	10	28.486	10.306	7	9.128	8

Data Base Values: $\mu = 11.623$, $\sigma^2 = 30.385$

Table 4

Average Ranking of PBIB and Random Sampling Plans

	$SE(\hat{\mu})$	Ψ_1	$SE(\hat{\sigma}^2)$	Ψ_2
PBIB	5.20	5.00	6.00	6.00
Simple Random Sampling	5.80	6.00	5.00	5.00

Summary and Conclusions

In this study, the use of PBIB designs to assign items to MMS subtests was compared with the results obtained when simple random sampling was used to build subtests. The conclusion is that PBIB designs are slightly superior for estimating the mean, but simple random sampling is slightly superior when estimating the variance.

The implication of these results for the educational evaluator is that PBIB designs can be used to build subtests, but the gain in precision is probably not worth the added effort required to use these designs. It is much easier to employ simple random sampling as a means of constructing subtests. While most evaluators probably are familiar with simple random sampling, the principles of PBIB designs, in all likelihood, would be unfamiliar to them.

References

Knapp, T. R. BIBD vs. PBIBD: an example of a priori item sampling. Unpublished manuscript, University of Rochester, 1968.

Knapp, T. R. Item sampling. Unpublished manuscript, University of Rochester, 1972.

Moy, M. L. Y., and Barcikowski, R. S. Item sampling: optimal number of people and items. The Journal of Experimental Education, 1974, 42, 46-52.

Shoemaker, D. M. Allocation of items and examinees in estimating a norm distribution by item-sampling. Journal of Educational Measurement, 1970 (a), 7, 123-128.

Shoemaker, D. M. Item-examinee sampling procedures and associated standard errors in estimating test parameters. Journal of Educational Measurement, 1970 (b), 7, 255-262.

Shoemaker, D. M. Further results on the standard errors of estimate associated with item-examinee sampling procedures. Journal of Educational Measurement, 1971, 8, 215-220.

Shoemaker, D. M. Standard errors of estimate in item-examinee sampling as a function of test reliability, variation in item difficulty indices and degree of skewness in the normative distribution. Educational and Psychological Measurement, 1972, 32, 705-714.

Shoemaker, D. M. A note on allocating items to subtests in multiple matrix sampling and approximating standard errors of estimate with the jackknife. Journal of Educational Measurement, 1973 (a), 10, 211-219.

Shoemaker, D. M. Principles and Procedures of Multiple Matrix Sampling. Cambridge: Ballinger Publishing Company, 1973 (b).